Timo Korkiakangas, Marco Passarotti

Challenges in Annotating Medieval Latin Charters

No annotation guidelines concerning substandard Latin are presently available. This paper describes an annotation style of substandard Latin that supplements the method designed for standard Latin by the Perseus Latin Dependency Treebank and the *Index Thomisticus* Treebank. Each word of the corpus can be assigned only one morphological analysis. In our system, the analysis can be either functional or formal. Functional analysis is applied when a form is language-evolutionarily deducible from the corresponding standard Latin form used in the same (semantico-)syntactic function (e.g. *solidus* pro *solidos* 'gold coins' as a direct object: analysis "accusative"). Formal analysis applies when no connection to the functionally required classical form exists (e.g. *heredibus* pro *heredes* 'heirs' as a subject: analysis "ablative" or "dative"). When running queries on the corpus, the formally analysed forms can be isolated, and percentages of standard and substandard forms can be counted. In addition, further principles concerning syntax and specific morphological issues are introduced.

1   Introduction

The present paper is related to a PhD project on the Latin case system in a corpus of ca. 500 Tuscan private charters (ca. 200,000 words) from the 8th and 9th centuries. So far, 1,452 sentences (28,488 words) have been annotated. Special attention is given to the core arguments (subjects and objects) and to prepositional phrases. The charters, published in three copyright-free diplomatic editions, have been digitised, proof-read and converted into XML.[1]

Research on the morphosyntax of the charters is performed by annotating the charters with the Latin Dependency Treebank (LDT) online tools provided by the Perseus Digital Library Project. The Latin and Ancient Greek Dependency Treebanks environment is suitable for our purpose, as it enables syntactic annotation, is user-friendly and publicly available.[2] Our annotation style is based on the *Guidelines for the Syntactic Annotation of Latin Treebanks* (BAMMAN et al. 2007[2]), which were launched to reconcile the practices of the annotators of LDT and the *Index Thomisticus* Treebank (IT-TB)[3] and to provide a general framework for all prospective treebanking projects in Latin. These guidelines and the related programs supporting annotation are designed for standard[4] Latin. The early medieval charters, however, differ from the standard in many respects (concerning orthography, morphology and syntax).

In this paper, we present a solution to the above-mentioned problem by introducing the concepts of formal and functional analysis plus further principles to supplement the existing guidelines. Even with these supplements, practical annotation requires highly subjective judgements on problematic cases, which is inevitable when dealing with charter texts and their language variety.

2   Standard and Substandard Latin

The *Guidelines for the Syntactic Annotation of Latin Treebanks* of LDT and IT-TB are designed according to the framework of dependency grammar as used on the analytical layer of annotation in the Prague Dependency Treebank (PDT) (HAJIČ et al. 1999) and adapted to Latin with the help of the *Latin Syntax and Semantics* of HARM PINKSTER (1990). Dependency grammar is an appropriate scheme of representation for highly inflected languages with a relatively free word order, such as Latin (BAMMAN et al. 2007[2], 3).

In LDT, both morphological and syntactic annotation is performed through a semi-automatic procedure provided by an online user interface. The morphological tagset reports information on the following: part of speech proper, person, number, tense, mood, voice, gender, case and degree. The syntactic annotation comprises syntactic tags (e.g. PRED, SBJ, OBJ, ATR, ADV) and head-dependent relations (BAMMAN et al. 2007[2], 4).[5]

If a word form already occurs in the treebank, the system provides its morphological analysis. If not, which is often the case when early medieval charters are concerned, the analysis must be typed manually in the table editor. If more than one analysis is provided by the system, annotators must choose the correct one from a drop-down menu. When combined, morphological and syntactic annotations allow performing advanced queries with *ad hoc* search engines, such as Annis used by LDT or Netgraph used by IT-TB.[6]

The Latin of the Italian charters from the 8th and 9th centuries is a technical, non-literary language variety resembling the style of the Lombard Laws. This variety seems to form a separate genre which is deliberately closer to the developments of spoken language than the literary texts of the same period, although it does not reflect spoken language directly nor is an attempt to act as a new "vulgar" language, distinct from Latin.[7]

The main issue concerning the Latin of early medieval charters is orthographic variation, which often concerns inflectional endings. These variations make it difficult to understand the syntactic structure of the texts. The existing annotation guidelines, designed for standard Latin, are not always able to manage substandard forms or standard forms used in a substandard way. Thus, new methods are needed with our corpus of medieval charters.[8] In standard Latin, each syntactic function is usually encoded by a relevant case form, which makes annotation process straightforward. However, with the Latin used in medieval charters, the equivalence between form and function is often not transparent.

3   The Solution: Functional and Formal Analyses

Each word in the corpus receives only one morphological tagging. In principle, we want to label all the forms functionally, i.e. according to their (semantico-)syntactic function in standard Latin. However, this is not always possible and several specifications are needed, mainly for nouns and other nominals.

If a word appears in its correct standard form, morphological tagging has no relevance since form and function are matching. If, however, a form is substandard, it is provided with a functionally based morphological analysis on condition that the form is language-evolutionarily deducible from the corresponding standard Latin form used in the same function. If no connection with the functionally required standard form exists, the substandard word is assigned a formal instead of a functional analysis.

For instance, we functionally annotate as accusatives the following substandard forms occurring as direct objects (although their form is not accusative) because they are meant to stand for the standard accusative forms: *solido* (standard: *solidum* 'gold coin'), *terra* (standard: *terram* 'land'), *testis* (standard: *testes* 'witnesses'), *solidus* (standard: *solidos* 'gold coins'). In *CDL* 23: *in tua cui supra emturi sit potestatem* (standard: *in tua cuius supra emptoris potestate* 'in the possession of you, the above purchaser'), the two words of the noun phrase *tua potestatem* ('your possession') are labelled functionally as singular ablatives dependent on the preposition *in*, although *potestatem* is formally an accusative singular in standard Latin. Finally, in *CDL* 45: *auris soledus trentas* (standard: *auri solidos triginta* '30 gold coins'), the standard ablative/dative plural form *auris* ('of gold') is functionally labelled as a genitive singular showing an additional *-s*.

Clear linguistic errors represent a class of their own and are always tagged according to their formal appearance. For instance, if a standard ablative/dative plural form, such as *heredibus* ('heirs'), functions as a subject (but does not occur in an ablative absolute construction), the form cannot be tagged functionally as a nominative because it is not possible to interpret it as a descendant of the nominative form. Thus, we label the *heredibus* according to its form, i.e. as ablative/dative plural. The form is an error probably due to the contamination between two or more formulae, a phenomenon common in medieval charters, or to the wrong interpretation of the abbreviation *hhd* (for *heredes*).

Sorting out such anomalous usages is relevant, as they are indirect (or "negative") clues of the corresponding, functionally correct form. This "negativity principle" represents, along with the functionality-formality approach, another pillar of our method. When running queries on the corpus, the distinction between formal and functional labelling allows us to isolate the formally analysed forms and to count the percentages of standard and substandard forms.

Both formal and functional labellings are based on standard Latin grammar. Although the language of these medieval charters may be quite different from standard Latin, analysing the charter texts in the framework of the traditional case system is justified because the charter texts try to resemble the standard language and, in spite of several disturbing factors, they reflect a multi-case system essentially similar to that of standard Latin. Adhering to standard Latin is also due to practical reasons: first, both LDT and IT-TB are based on standard Latin grammar; second, the language of the utilitarian texts of the 8th and 9th centuries, such as charters and laws, was never described in terms of prescriptive grammar similar to that of Classical Latin.

Distinguishing between functional and formal analyses is not the only possible method for annotating substandard Latin. In principle, one could also provide both types of annotations side by side, but this sort of multilevel annotation would be often redundant, as it would reduplicate the same information in most cases.

Another possible solution would be to provide functional analysis only, thus refining the query results according to the endings (for instance, by selecting all the subjects ending in *-ibus*). However, this solution would result in clearly erroneous analyses: for instance, the form *heredibus* would be tagged as "nominative". Our purpose is to provide morphological analyses that reflect the real language-evolutionary origin of the forms, in order to make

possible both to exploit the morphological tagging and to detect the 'anomalous' cases, i.e. those whose morphological tagging is incompatible with their syntactic function (reported by the dependency relation tag).

4 Additions to the LDT/IT-TB Guidelines

The principles described in the previous section are the backbone of our annotation style. This section introduces further specifications and individual rules designed in order to treat recurrent problematic structures consistently. This is of special relevance to morphology because it differs extensively from standard Latin.

4.1 Lemmatisation

Reducing lemmas. Almost all the words in the charters have two or more graphical variants. Likewise, one single morph may have several realisations. Therefore, particular attention must be paid to lemmatising all its graphical variations under one common lemma in order to avoid proliferation of lemmas in the Perseus Dynamic Lexicon database (BAMMAN – CRANE 2008, 11–13). For instance, nouns facing gender change, such as the masculine nominative plural *saeculi* ('centuries'), as well as adjectives facing declension change, such as the second declension nominative singular *inanus* ('void'), are lemmatised under the standard lemmas: *saeculum* (neuter) and *inanis* (third declension), respectively. The aim is to respect the choices taken by the scribe as far as they are traceable. This is also the motive for formally labelling those functionally impossible case forms, such as *heredibus*, in order to show their anomalous status.

Proper names. Several Germanic and Latin proper names exhibit much variation. For example, the form *Delmati* is lemmatised under *Dalmatius* and the forms *Guntifrido* and *Cuntefrid* under *Guntifridus*. However, it is sometimes difficult to establish the correct lemma, as no variant seems to be more justified (or more frequent) than the others. In the charters, there are also several unidentified place names. Unknown second declension toponyms, such as *Brancalo*, are lemmatised as neuters ending in -*um*. Although in some cases the lemma can be reconstructed on the grounds of the modern name of the place in question[9], those names that are completely opaque must be labelled as "unknown".

4.2 Syntax

Omitted elements. As our research focuses primarily on the syntactic constructions concerning the core arguments (subjects and direct/indirect objects) and prepositional phrases, we leave unannotated all the non-nominal adverbials, except negation particles, and the punctuation marks, except those commas which have a role in coordinated or appositive tree structures (cf. the lacking "," and "*et*" in Figure 1). Terminal punctuation marks are always tagged with the technical label AuxK (BAMMAN et al. 2007[2], 33–34).
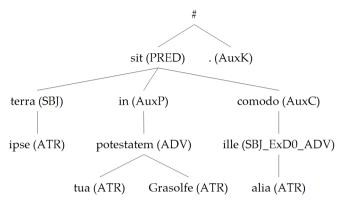
Figure 1: Dependency tree of *CDL* 220: *ipse terra in tua Grasolfe sit potestatem, comodo et ille alia.*

Ellipsis and fragmentary parts. In annotating ellipsis, we follow the LDT style and reconstruct the omitted nodes. In fact, the formulaicity of the charters very often allows deducing even the exact wordings of the missing parts with a high degree of reliability. For example, in *CDL* 220: *ipse terra in tua Grasolfe sit potestatem, comodo et ille alia* ('this plot be in your possession, Grasolfus, just as that other one'), the omitted verb of the subordinate clause *comodo* (*sit*) *et ille alia* ('just as (be) that other one') is reconstructed through the complex tag SBJ_ExD0_ADV. This means that *ille* ('that') is the subject (SBJ) of the omitted verb (ExD0, "externally dependent", is the technical label for missing items) that in the tree would be the head of an adverbial (ADV) subordinate clause (see Figure 1). This is the only aspect where the annotation style of IT-TB differs from the LDT one. As a matter of fact, IT-TB (as PDT) does not resolve the ellipsis and, thus, would assign to *ille* the simple tag ExD. In those cases where an elliptic structure is ambiguous or where words are missing because the original source is damaged, we follow the IT-TB style and link the orphan nodes directly to their assumed parents via ExD (BAMMAN et al. 2007[2], 36–37; BAMMAN et al. 2007[1], 4).

Indirect objects. We introduce a specific tag (c="1") to annotate indirect objects while LDT and IT-TB use the same label OBJ for both direct and indirect objects. Even though the latter solution is suitable for standard Latin, where indirect objects always occur in dative or as prepositional phrases, it cannot be applied to our texts, which feature a high degree of formal variation. In *CDL* 125: *in terra, que offerui sancti Petri cum ipsa fossa* ('in the plot, which we donated to St. Peter, with the ditch'), the direct object is *que* (standard: *quam*) and the indirect object is *sancti Petri* (standard: *sancto Petro*). Although they are both labelled with OBJ (see Figure 2), *sancti Petri* is assigned the additional tag c="1" in order to make clear its status as an indirect object.[10] In this case, the morphological formal analysis of *sancti Petri* (genitive singular) also helps to detect the anomaly.

Vocatives. Although the Guidelines demand to link the vocatives to their verbal heads with the label ExD (BAMMAN et al. 2007[2], 41), we link them to their nominal heads via ATR since, in our charters, the vocatives mainly represent the function of nominal attributives.

See, for example, the words *Uuarniperte* and *Lamprande* in *CDL* 269: *uouis Uuarniperte et Lamprande presbiteri* ('to you, priests Warnipertus and Lamprandus'), and *Grasolfe* in Figure 1: *in tua Grasolfe sit potestatem* ('in your possession, Grasolfus').
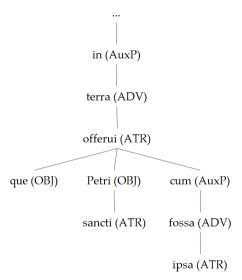
```
                              ...
                               |
                          in (AuxP)
                               |
                         terra (ADV)
                               |
                         offerui (ATR)
                        /      |      \
              que (OBJ)    Petri (OBJ)    cum (AuxP)
                               |              |
                          sancti (ATR)    fossa (ADV)
                                              |
                                          ipsa (ATR)
```

Figure 2: Dependency tree of *CDL* 125: *in terra, que offerui sancti Petri cum ipsa fossa.*

4.3 Morphology

Subjects. The following annotation style only applies to subjects of clauses whose verb occurs in finite form. The standard case of a subject headed by a finite verb is nominative. The subjects of *accusativus cum infinitivo* constructions and ablative absolutes are not discussed here. In standard Latin, they are encoded with accusative and ablative, respectively.

The second and fourth declension masculine singular subjects ending in <-o -u -um>, such as the second declension form *Deo* (standard: *Deum* 'God'), are tagged formally as accusatives because, according to the bicasual hypothesis, they cannot be deduced from the standard nominative form. The neuter subjects ending in <-o -u -um>, such as *pretio* (standard: *pretium* 'price'), are tagged functionally as nominatives. In principle, the neuter subjects could equally well be tagged as accusatives because the standard nominative and accusative forms of the second declension neuters are identical.

The formal tagging also applies to those third declension singular subjects ending in <-e -em> whose stem has an additional syllable in all cases except nominative, such as nominative *potes-tas* vs. accusative *potes-ta-tem* ('possession'). Instead, those third declension singular subjects ending in <-e -i -em> whose stem has the same number of syllables in all cases, such as nominative *tes-tis* vs. accusative *tes-tem* ('witness'), and all the first declension singular subjects ending in <-a -am> are tagged functionally as

nominatives because the word-final /*m*/ was no more pronounced in Late Latin. For example, *potestate* and *potestatem* as subjects are tagged as accusatives, while *teste* and *testem* as subjects are tagged as nominatives.

These principles are based on the following reason: When annotating third declension subjects with an equal number of syllables in all cases, no distinction can be made between the language-evolutionary outcomes of the standard nominative and accusative forms. However, in the second and fourth declension subjects as well as in the third declension subjects with an additional stem syllable, the nominative and the accusative forms still differed from each other because the second and fourth declension final /*s*/ in the nominative and the third declension stem extension mark the nominative as distinct. Indeed, the opposition between the second declension nominative singular and accusative forms seems to have been partially neutralised in the Latin of Tuscany in the 8th and 9th centuries, but the bicasual assumption is a good working hypothesis (cf. ZAMBONI 2000, 233–235, 243–244). Cases such as the second declension subject *Deo* (tagged as an accusative) illustrate the role of subjective decisions in the annotation process: determining the status of certain forms is not uncontroversial, but the decisions can be systematic if they are based on a well-grounded theory. Indeed, the annotation style depends considerably on the chosen theoretical framework, and the choice of the annotation framework is dictated by the purpose of the corpus. As our corpus is mainly designed for studying case marking, the bicasual assumption seems to be a valid background assumption for our annotation style.

In the fifth declension, the confusion appears to be so massive that the analyses must be especially delicate. In the plural forms of all the declensions, the deviations from the standard forms are fewer.

Genitives and the oblique case. In late substandard Latin, the genitive was often replaced by an oblique case, most likely derived from the standard accusative. The development, which started in spoken language, led to a situation where the standard case system was probably reduced to a bicasual system (nominative vs. accusative). The accusative gradually absorbed the functions of all the other cases except the nominative and finally even that of the nominative (VÄÄNÄNEN 1981, 116–117; cf. ZAMBONI 2000, 248). The oblique forms are tagged formally as accusatives. For instance, in the subscription *CDL* 261: *signum manus Alprand filio quondam Teuduald testis* ('mark of the hand of Alprandus, son of late Teudualdus, witness'), the word *filio* ('son') is labelled as an accusative and linked to its head *Alprand* via ATR.

Prepositions. As a general principle, we label as accusatives the complements of prepositions governing accusative in standard Latin and as ablatives the complements of prepositions governing ablative, if the case-endings can be claimed to represent the original accusative and ablative forms, respectively. This requires looking at the meanings of the prepositional phrases, as some prepositions govern different cases according to what they mean. For instance, the prepositions *in* and *super* govern accusative when expressing motion and ablative when expressing state. In *CDL* 23: *sup die quartum* ('on the fourth day'), we label both *die* ('day') and *quartum* ('fourth') as singular ablatives since *sub* governs ablative if it means state, and accusative if it means motion.

Nominal attributives. Nominal attributives occur mainly in the titles of commissioners and addressees of legal transactions, for example in *CDL* 266: *ego Autulu uir religiosus clirico filio quondam Bonuald de uico Turrite* ('I Autulus, *vir religiosus*, clerk and son of late Bonualdus from the village of Turrite'). Several problems arise when the head-dependent relations in such noun phrases are labelled. As a rule, we choose as the head of the noun phrase the member with the highest ranking in the following hierarchy of animacy: personal pronouns > proper names > other nouns referring to humans. Thus, the head of the above noun phrase is *ego*, under which *Autulu* is attached as an attributive; *uir religiosus*, *clirico* and *filio* are then linked to *Autulu* as attributives.

Absolute constructions. Some substandard absolute constructions, such as the accusative absolute and the *post* construction, had been quite firmly established even in the late literary language (HELTTULA 1987, 6–7, 91–92). As far as morphological annotation is concerned, we do not force the absolute structures into the form of standard Latin. In the medieval charters, almost all case forms can occur in absolute constructions, and we do not want to reduce such formal variety to any expected pattern, such as accusative absolute, because we take into account the scribes' freedom in choosing the case form in absolute constructions.

This applies, for instance, when a case form might be interpreted as a descendant of the standard ablative. For example, in *CDL* App. *postea*, *inimicum eum suadente* (standard: *inimico eum suadente*), *inuolauit mihi ipsam cartulam* ('later on, he stole me the charter, incited by the Devil'), the noun *inimicum* ('Devil') can be interpreted as an ablative, but the structure rather seems to be an accusative absolute.

Post constructions are treated as if they were normal prepositional constructions. Examples of these are *MED* 424: *post fructum de ipsa res recollecto* ('having collected the yield on this property') and *CDL* 260: *spondeo … conponere tibi post hanc cartulam ostensam … quae tibi subtraxerimus* ('I promise … to compensate you for what we may have seized from you, if this charter is brought in evidence') (see Figure 3).

Vocatives. The label "vocative" is assigned only to the forms showing a clear vocative ending, such as *Uuarniperte* and *Lamprande* in the above-mentioned *CDL* 269: *uouis Uuarniperte et Lamprande presbiteri*; the form *presbiteri* is tagged as a nominative plural.

Gender change. Gender change from neuter singular to masculine singular and from neuter plural to feminine singular is a relevant example of the changes occurring in Latin declension. In our annotation style, the neuters occurring in masculine or feminine forms are lemmatised under their standard lemmas and still labelled as neuters. For instance, see *pretius* ('price', masculine) in *CDL* 66: *suscipemus … pretius* (standard: *suscepimus … pretium* 'we received … the price', neuter), or *adiacentia* ('neighbourhood', feminine) in *CDL* 266: *cum omnem adiacentia sua ... pertenente* (standard: *cum omnibus adiacentibus suis … pertinentibus* 'with all its neighbourhood … that belongs to…', neuter). Thus, the annotation does not reveal gender change. This is only revealed when the words labelled as neuters are sorted by their endings or when they are read in their context, as in *cum omnem adiacentia sua ... pertenente*.
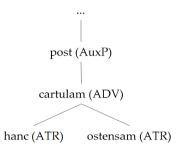
...
|
post (AuxP)
|
cartulam (ADV)

hanc (ATR)   ostensam (ATR)

Figure 3: Dependency tree of *CDL* 260: *post hanc cartulam ostensam.*

Number and person in verbs. If it is not possible to determine the number of a verb, it is tagged according to its formal appearance. This phenomenon mainly occurs with the third person of verbs expressing actions performed by the addressee of a charter, as it is often unclear whether the addressee is acting alone or with his/her heirs. However, in *CDL* 23: *petras que iniui esse uiditor* ('the stones that are (lit. is) seen there') the singular verb *uiditor* may be due to impersonalisation of the passive structure (GIANOLLO 2005, 100). The relative pronouns (*que* 'that') were already on their way to becoming indeclinable.

The person of the verb is usually tagged functionally because the person is normally easier to recognise than the number. The context may be helpful: for instance, in *CDL* 28: *abbas … habeas* ('the abbot … may have'), the form *habeas* 'you may have' is analysed as a third person singular (standard: *habeat*). In more complex cases, such as *donatores … habeas* (standard: *habeant* 'the donators … may have'), the annotator has to make delicate decisions which depend on the amount of graphical variation observed in the charter.

5   Two Case Studies

In order to demonstrate how helpful the distinction between formal and functional analysis is for organising and retrieving data, we briefly report two case studies concerning two simple constructions which occur in our corpus.

The first construction concerns those prepositional phrases that are headed by the preposition *ad* ('to'). In Latin, the preposition *ad* governs nouns and pronouns inflected in the accusative case. In our corpus, however, several exceptions to this rule occur: these exceptions can be retrieved by exploiting the annotation. Table 1 reports the results concerning this construction.

	Non-accusative case	Accusative case
Formal = Functional	---	81
Functional	---	119
Formal	19	---
TOTAL by case	19	200

TOTAL	219

Table 1. The results concerning prepositional phrases headed by *ad*.

In the part of our corpus annotated so far, there are 219 lexical items governed by the preposition *ad*. Among them, only 19 are tagged formally as clearly non-accusative forms showing no connection with the functionally required standard form. We report two examples of such items, one presenting a direct governance and the other showing a coordinated governance: *ad heredibus uestris* ('to your heirs') and *ad Laurentio et Ualentini* ('to Laurentius and Valentinus').

The remaining 200 items are tagged as accusatives. Among them, 81 are standard accusative forms (*ad ecclesiam*: 'to the church') and, thus, their formal and functional tagging are matching. The remaining 119 items are substandard accusative forms (language-evolutionarily deducible from the corresponding standard Latin forms): hence, they are tagged functionally as accusatives (*ad ecclesia*).

The second construction is ablative absolute. This construction consists of one participle (in the ablative case) and one subject (also in the ablative case). Table 2 presents the results concerning this construction.

	Subject in non-ablative case	Subject in ablative case
Formal = Functional	---	47
Functional	---	71
Formal	18	---
TOTAL by case	18	118
TOTAL	136	

Table 2. The results concerning ablative absolute.

On a total of 136 items occurring as subjects of ablative absolute constructions, only 18 are annotated formally as clearly non-ablative forms. One example is *Dominus interueniente* ('with the intervention of God'), where *Dominus* is a nominative form. Most of the ablative absolute constructions present a subject tagged as ablative (118 cases). Among these, 47 are tagged as standard ablative forms (*regnante Liutprando*: 'under the reign of Liutprand'); 71 are substandard ablative forms, which are tagged functionally (*regnante Liutprando*).

6   Conclusions

In order to overcome the incompatibility between the annotation of Latin in the medieval charters and the annotation style provided by the LDT/IT-TB guidelines, two distinct forms of analysis (formal and functional) and a number of additional principles were introduced.

Four issues can be distinguished: (a) functional analysis is applied when a form is deducible from the corresponding standard Latin form used in the same function; (b) formal analysis is applied when a form is not deducible from the standard Latin form used in the same function; (c) the linguistically impossible forms can be isolated when querying the

data; (d) the query results of the data can be further processed by classifying the results according to endings; the percentages of standard, early medieval and linguistically impossible forms can be counted.

Building and querying an annotated corpus of substandard language that shows much variation is a challenging task. An inherent disadvantage of introducing new rules in annotation is that the corpus becomes more difficult to use. The user must consider several different parameters that were applied when building the annotated corpus. This, along with separating formal and functional labellings, implies that the pure quantitative results from the queries on our corpus cannot be compared with those acquired from corpora in standard Latin. However, following the same general principles of syntactic annotation (in terms of theoretical framework, syntactic labels and head-dependent attachment) allows us to compare the syntactic constructions occurring in our corpus with those of LDT and IT-TB.

Acknowledgements

---

[1] The three editions are *Codice diplomatico longobardo* (*CDL*) 1–2 (LUIGI SCHIAPARELLI, 1929–1933); *Codice diplomatico toscano*, part 2, vol. 1 (FILIPPO BRUNETTI, 1833) and *Memorie e documenti per servire all'istoria del Ducato di Lucca* (*MED*), part 5, vol. 2 (DOMENICO BARSOCCHINI, 1837). *CDL* is digitised and proof-read by the Institut für Mittelalterforschung of the Austrian Academy of Sciences while the other two are digitised by Google and proof-read by us. Almost all the charters were also published recently in the *Chartae Latinae Antiquiores* (2nd series).

[2] The Perseus Latin and Ancient Greek Dependency Treebanks are projects aimed at treebanking texts in Classical Latin and Greek; they are both hosted at Tufts University in Boston, USA (*http://nlp.perseus.tufts.edu/syntax/treebank/index.html*). Another project in the field is the Laboratoire d'Analyse Statistique des Langues Anciennes in Liège, Belgium (LASLA, *http://www.cipl.ulg.ac.be/Lasla/*). The annotation style of syntax by LASLA concerns subordination patterns only.

[3] The *Index Thomisticus* Treebank is an ongoing project aimed at the syntactic annotation of the *Index Thomisticus*, a morphologically annotated corpus of the texts of St. Thomas Aquinas. The project is hosted at the Catholic University of the Sacred Heart in Milan, Italy (*http://itreebank.marginalia.it).*

[4] By "standard" Latin we mean the variant of Latin mostly used by the Classical authors and reported in the pedagogical grammatical tradition.

[5] For the morphological tagset, see the README file for the Latin Dependency Treebank at *http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/README.txt*.

[6] The Annis search engine is not yet publicly available. The *Index Thomisticus* Treebank can be browsed through Netgraph at *http://gircse.marginalia.it/~passarotti/netgraph/client/applet/NGClientAen.html*.

[7] See BARTOLI LANGELI 2006, 24–28 about the status of the Latin of the Lombard charters and laws.

[8] See PHILIPPART DE FOY [forthcoming] about changes in the LASLA annotation procedures to face similar problems in a medieval hagiographic corpus.

[9] The *Chartae Latinae Antiquiores* editions usually report the modern equivalents of the place names occurring in the charters.

[10] The verb heading a relative clause is linked to its antecedent as an attributive (ATR) (BAMMAN et al. 2007[2], 37–38).

References

BAMMAN, D. – CRANE, G. (2008). „Building a Dynamic Lexicon from a Digital Library". In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008, Pittsburgh). New York: ACM, 11–20.

BAMMAN, D. – PASSAROTTI, M. – CRANE, G. – RAYNAUD, S. (2007[1]). „A Collaborative Model of Treebank Development". In: Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT2007, Bergen), 1–6.

BAMMAN, D. – PASSAROTTI, M. – CRANE, G. – RAYNAUD, S. (2007[2]). Guidelines for the Syntactic Annotation of Latin Treebanks (v. 1.3). *http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/ guidelines.pdf*

BARTOLI LANGELI, A. (2006). Notai. Scrivere documenti nell'Italia medievale. Roma: Viella.

CDL = Codice diplomatico longobardo 1–2. A cura di LUIGI SCHIAPARELLI. (1929–1933). Roma: Tipografia del Senato.

GIANOLLO, C. (2005). „Middle Voice in Latin and the Phenomenon of Split Intransitivity". In: Calboli, G. (ed.) (2005). Latina lingua! Proceedings of the Twelfth International Colloquium on Latin Linguistics (ICLL 2003, Bologna). Roma: Herder, 1: 97–110.

HAJIČ, J. – PANEVOVÁ, J. – BURÁŇOVÁ, E. – UREŠOVÁ, Z. – BÉMOVÁ, A. (1999). Annotations at Analytical Level. Instructions for annotators. Institute of Formal and Applied Linguistics, Prague. *http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/aman_en.pdf*

HELTTULA, A. (1987). Studies on the Latin Accusative Absolute. Tammisaari: Societas Scientiarum Fennica.

MED = Memorie e documenti per servire all'istoria del Ducato di Lucca 5:2. A cura di DOMENICO BARSOCCHINI. (1837). Lucca: Francesco Bertini.

PHILIPPART DE FOY, C. [forthcoming] „Lemmatiser un corpus de textes hagiographiques: enjeux et modalités pratiques". In: Biville, F. (ed.) Latin vulgaire – latin tardif IX. Actes du IXe colloque international sur le latin vulgaire et tardif (LVLT 2009, Lyon).

PINKSTER, H. (1990). Latin Syntax and Semantics. London: Routledge.

VÄÄNÄNEN, V. (1981). Introduction au latin vulgaire. Paris: Éditions Klincksieck: (Bibliothèque française et romane, A:6.

ZAMBONI, A. (2000). „L'emergere dell'italiano: per un bilancio aggiornato". In: Herman, J. – Marinetti, A. (eds.) (2000). La preistoria dell'italiano. Atti della Tavola Rotonda di Linguistica Storica. Università Ca' Foscari di Venezia, 1999. Tübingen: Niemeyer, 231–260.